下载APP

14｜留存分析：哪些因素会影响用户的留存率？

2021-09-29 黄佳

《零基础实战机器学习》 课程介绍 >

讲述：黄佳

时长 20:24 大小 18.70M

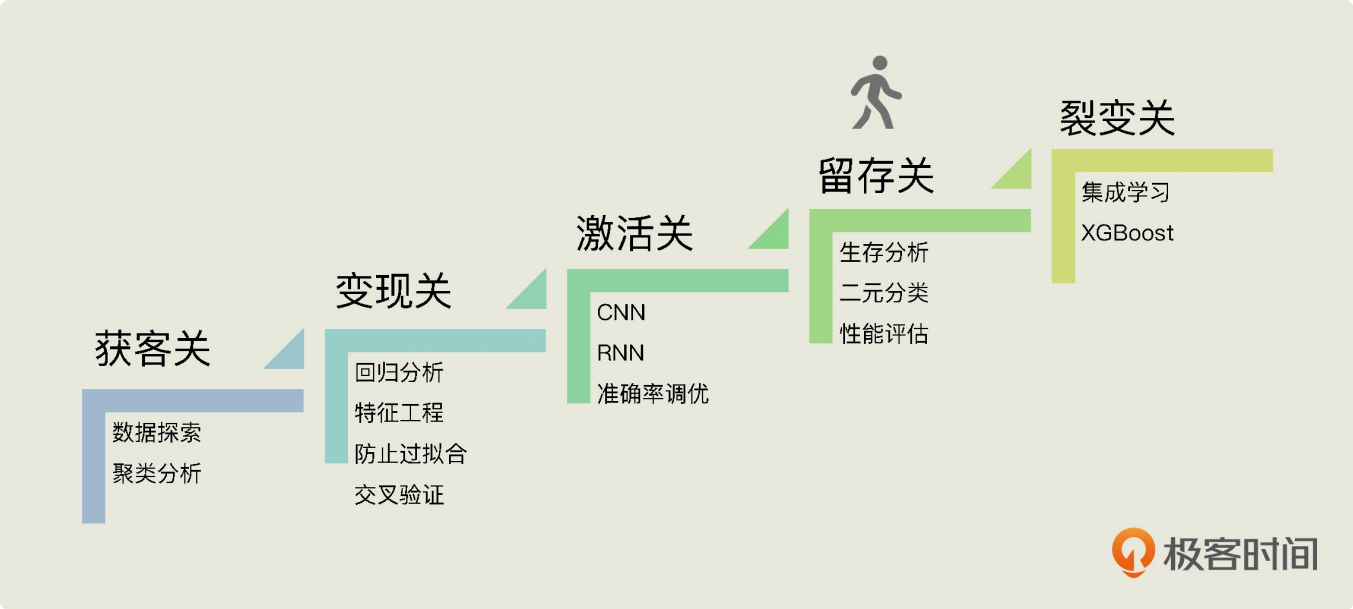
你好，我是黄佳。欢迎你来到零基础实战机器学习，今天我们正式开启留存关。

在开启这一关之前，我想给你讲一个小时候的故事。这是 70 年代人的回忆了，那时候，小学的校门口有两家租“小人书”摊，一毛钱看一本。小人书，就是一大堆的 64 开小画本，里面是西游记、水浒传、封神榜、射雕英雄传啥的。

两家的小人书其实都差不多的，那个年代也没有玄幻，也没有网络写手。可是呢，A 家的生意总比 B 家好，我一开始不知道怎么回事，后来我想起 A 家的老板记忆力很好，他总能记住我们这些孩子谁是谁，而且他每次都和我们说：“看五本送一本哈！”如果我们看三本，他就会说：“下回再看两本就多送一本免费的。”

多年以后，我恍然大悟，原来老板在 30 年前就建立起了买 5 送一和老客户会员卡的制度。只不过他是凭借自己的记忆力来完成的。正是由于这个原因，A 老板的用户留存率比较高。

什么是用户留存？留存就是让老用户一直使用你的产品。其实，谈留存，就不能不提另外一个关键概念“流失”。用户的留存和流失，是阴和阳的两极，此消彼长。收集和分析用户行为数据，有助于我们摸清用户是“留”还是“失”之间的微妙门道。



那么，今天这节课我们就通过机器学习模型来显示易速鲜花的会员留存曲线，并且分析“易速鲜花”会员卡中设置的哪些机制，是对会员留存影响比较大的因子，哪些又容易造成用户的流失。学习了这些，你就掌握了一组非常实用的运营分析工具，这对于如何留客有很大的指导意义。

定义问题

现在，让我们回到易速鲜花公司的运营部，一起为提升“易速鲜花”的用户留存率而奋斗。易速鲜花其实已经建立起了一套完善的会员机制，来增加用户的黏性。不过，运营部门的老总不确定的是，在一系列的促销套餐和会员卡类型中，哪些因素和用户的留存关系更为密切。

你从运营那边拿到了这样的数据集：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	用户码	性别	玫瑰套餐	紫罗兰套	郁金香套	百合套餐	康乃馨套	胡姬花套	生日套餐	情人节套	会员卡类	入会月数	会费支付	平均月消	总消费	已停付会费
2	1	女	是	是	否	是	否	是	是	否	年卡	9	手工转账	65.6	593.3	否
3	2	男	否	否	否	否	否	否	否	是	月卡	9	手工转账	59.9	542.4	否
4	3	男	否	否	否	否	是	否	否	否	月卡	4	就餐时付	73.9	280.85	是
5	4	男	是	否	否	是	是	否	是	是	月卡	13	就餐时付	98	1237.85	是
6	5	女	否	是	否	否	否	是	是	是	月卡	9	花呗付款	69.4	571.45	否
7	6	女	是	否	是	是	是	是	是	是	双年卡	71	微信自动	109.7	7904.25	否
8	7	女	是	否	否	否	否	是	是	否	月卡	3	手工转账	83.9	267.4	是
9	8	男	是	否	是	否	否	是	否	否	双年卡	63	花呗付款	84.65	5377.8	否
10	9	女	否	否	是	否	否	否	否	否	月卡	7	就餐时付	48.2	340.35	否
11	10	女	是	是	是	是	是	是	是	是	双年卡	65	手工转账	90.45	5957.9	否
12	11	男	否	否	是	否	否	是	是	否	双年卡	54	花呗付款	45.2	2460.55	否
13	12	女	是	是	是	是	是	是	是	是	双年卡	72	花呗付款	116.8	8456.75	否
14	13	女	是	否	否	否	否	否	否	否	月卡	5	就餐时付	68.95	351.5	否
15	14	女	否	否	是	是	是	否	是	否	双年卡	72	微信自动	101.3	7261.25	否
16	15	女	否	否	是	是	是	是	否	否	双年卡	56	微信自动	45.05	2560.1	否
17	16	女	是	是	否	是	是	否	否	是	双年卡	71	花呗付款	95.75	6849.4	否

在这个数据集中，有下面一系列的字段：

用户码；

性别；

一系列的会员专享套餐优惠，包括玫瑰套餐、紫罗兰套餐、郁金香套餐、百合套餐、康乃馨套餐、胡姬花套餐、生日套餐、情人节套餐等，这个是注册会员时，随机赠送给会员的；

会员卡的类型；

入会月数 ；

会费支付方式；

平均月消费；

总消费；

是否已停付会费。

我们今天的目标呢，就是从这个数据中找出，哪些因素和用户的留存关系更为密切，哪些因素可能会导致老用户的流失。而这个问题，既不是一个回归问题，也不是一个分类问题，它是一个数据探索类型的问题。

对于这个问题，首先，我们要做的仍然是数据集的导入，数据清洗，以及简单的可视化工作。

数据的预处理

1. 数据导入

数据集我们已经有了，你在这里可以下载到。现在，第一步是导入必须的库，并载入数据集。

复制代码

```
import numpy as np #导入NumPy
import pandas as pd #导入Pandas
df_member = pd.read_csv('易速鲜花会员留存.csv') #载入数据集
df_member.head() #显示数据头几行
```

输出如下：

	用户码	性别	玫瑰套餐	紫罗兰套餐	郁金香套餐	百合套餐	康乃馨套餐	胡姬花套餐	生日套餐	情人节套餐	会员卡类型	入会月数	会费支付方式	平均月消费	总消费	已停付会费
0	1	女	是	是	否	是	否	是	是	否	年卡	9	手工转账	65.60	593.3	否
1	2	男	否	否	否	否	否	否	否	是	月卡	9	手工转账	59.90	542.4	否
2	3	男	否	否	否	否	是	否	否	否	月卡	4	就餐时付费	73.90	280.85	是
3	4	男	是	否	否	是	是	否	是	是	月卡	13	就餐时付费	98.00	1237.85	是
4	5	女	否	是	否	否	否	是	是	是	月卡	9	花呗付款	69.40	571.45	否
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	7039	女	否	否	是	否	否	是	否	否	年卡	13	手工转账	55.15	742.9	否
7039	7040	男	否	否	否	是	否	否	否	否	月卡	2	手工转账	50.30	92.75	否
7040	7041	男	是	否	否	否	否	否	否	是	月卡	22	就餐时付费	85.10	1873.7	是
7041	7042	男	是	是	是	否	是	是	否	是	双年卡	67	手工转账	67.85	4627.65	否
7042	7043	男	是	是	是	是	是	否	是	是	双年卡	63	就餐时付费	59.00	3707.6	否

7043 rows × 16 columns

从输出中，我们看到这个数据集一共有 7043 条易速鲜花会员的数据。下面我们就开始做数据清洗的工作。

2. 数据清洗

我们先用 describe 方法看一看这个数据集。

复制代码

```
df_member.describe() # 显示数据集数值字段概率
```

输出如下：

	用户码	入会月数	平均月消费
count	7043.000000	7043.000000	7043.000000
mean	3522.000000	32.371149	64.761692
std	2033.283305	24.559481	30.090047
min	1.000000	0.000000	18.250000
25%	1761.500000	9.000000	35.500000
50%	3522.000000	29.000000	70.350000
75%	5282.500000	55.000000	89.850000
max	7043.000000	72.000000	118.750000

不知道你有没有观察到一个有意思的事儿，就是这个 describe 方法把用户码当成了数值性数据（这也难免，因为这里的用户码是数字编号）。此外，这个表格中没有“总消费”这个字段。我推测“总消费”这个字段有点问题，它有可能看起来是数值，但是实际的格式是字符串。

我用 is_numeric_dtype 方法验证一下我的推测，这个方法会显示一个字段是否为数值类型。

复制代码

```
from pandas.api.types import is_numeric_dtype #导入is_numeric_dtype工具
is_numeric_dtype(df_member['总消费']) #是否为数值字段？
```

输出显示：

复制代码

```
False
```

果然，“总消费”这个字段的格式有问题，目前它并不是一个数值字段，只是看起来像是数值的字符串。这里我们用 to_numeric 这个 API 对它做个转换：

复制代码

```
df_member['总消费'] = pd.to_numeric(df_member['总消费'], errors='coerce') #把总消
df_member['总消费'].fillna(0, inplace=True) #补充0值
```

然后，我们再用 is_numeric_dtype 验证，就会得到输出结果 True，这说明“总消费”字段的格式成功转换成了“数值型”。

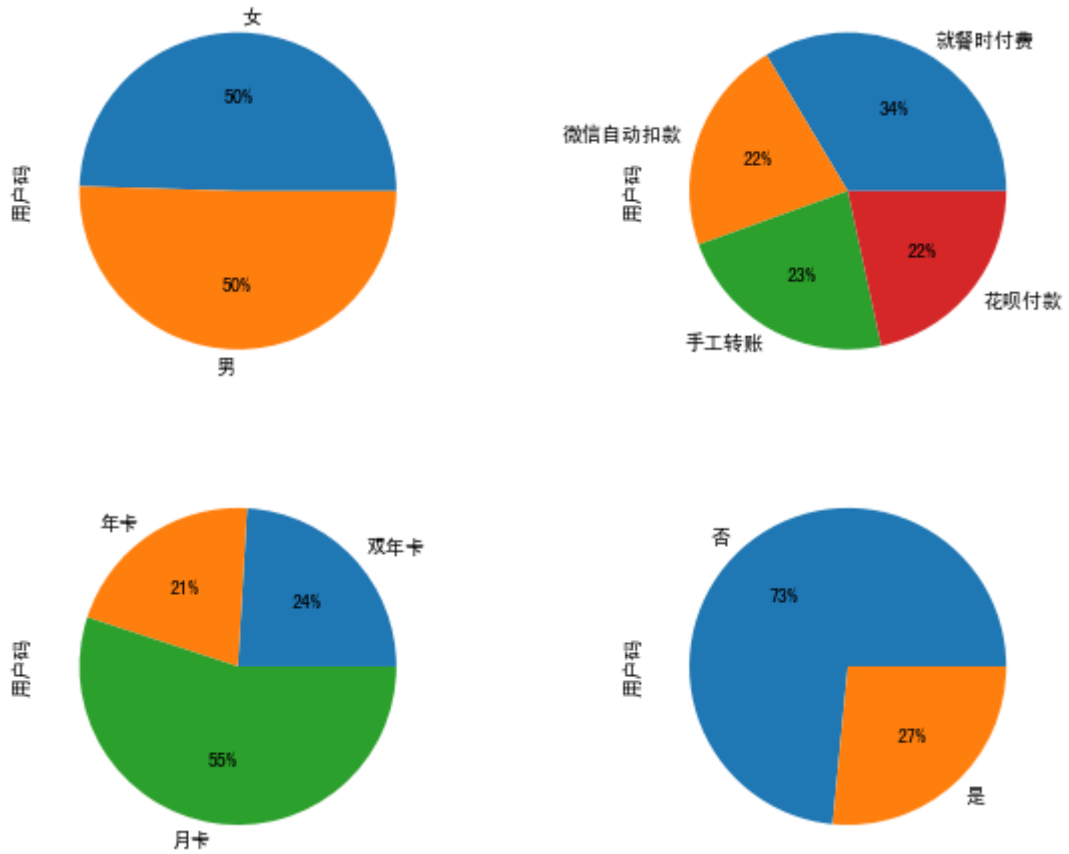
3. 数据可视化

在数据可视化这个部分，我们通过饼图来看一看性别、会费支付方式、会员卡类型、已停付会费这四个字段的分布比例。这些分布情况，尤其是留存与流失的会员占比，对于我们确定当前运营的重点关注点很有指导意义。

复制代码

```
import matplotlib.pyplot as plt #导入绘图工具
plt.figure(figsize=(10,8)) #图片大小
plt.subplot(2, 2, 1) #子图1
ax = df_member.groupby('性别').count()['用户码'].plot.pie(autopct='%1.0f%%') #饼
plt.subplot(2, 2, 2) #子图2
ax = df_member.groupby('会费支付方式').count()['用户码'].plot.pie(autopct='%1.0f%
plt.subplot(2, 2, 3) #子图3
ax = df_member.groupby('会员卡类型').count()['用户码'].plot.pie(autopct='%1.0f%%'
plt.subplot(2, 2, 4) #子图4
ax = df_member.groupby('已停付会费').count()['用户码'].plot.pie(autopct='%1.0f%%'
plt.show() #显示
```

输出如下：



当然，这些信息，其实你也可以通过柱状图来显示。在上面这张图中，我希望你注意一下右下角“已停付会费”这个分布图，因为它能直接体现出我们的用户是否流失，而用户是否流失就是我们的标签。

下面，我们开始做一些特征工程，以确保所有字段都能被机器学习模型处理。

4. 特征工程

我们再来看看这个数据集：

	用户码	性别	玫瑰套餐	紫罗兰套餐	郁金香套餐	百合套餐	康乃馨套餐	胡姬花套餐	生日套餐	情人节套餐	会员卡类型	入会月数	会费支付方式	平均月消费	总消费	已停付会费
0	1	女	是	是	否	是	否	是	是	否	年卡	9	手工转账	65.60	593.3	否
1	2	男	否	否	否	否	否	否	否	是	月卡	9	手工转账	59.90	542.4	否
2	3	男	否	否	否	否	是	否	否	否	月卡	4	就餐时付费	73.90	280.85	是
3	4	男	是	否	否	是	是	否	是	是	月卡	13	就餐时付费	98.00	1237.85	是
4	5	女	否	是	否	否	否	是	是	是	月卡	9	花呗付款	69.40	571.45	否
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	7039	女	否	否	是	否	否	是	否	否	年卡	13	手工转账	55.15	742.9	否
7039	7040	男	否	否	否	是	否	否	否	否	月卡	2	手工转账	50.30	92.75	否
7040	7041	男	是	否	否	否	否	否	否	是	月卡	22	就餐时付费	85.10	1873.7	是
7041	7042	男	是	是	是	否	是	是	否	是	双年卡	67	手工转账	67.85	4627.65	否
7042	7043	男	是	是	是	是	是	否	是	是	双年卡	63	就餐时付费	59.00	3707.6	否

7043 rows × 16 columns

在这个数据集中，最后一列“已停付会费”字段显示有一些用户已流失。目前这个字段的值的类型是汉字，我们需要转换为 0、1 值才能被模型读入。同理，我们对“性别”字段也做类似处理。

复制代码

```
# 把汉字转换成0、1值
df_member['已停付会费'].replace(to_replace='是', value=1, inplace=True)
df_member['已停付会费'].replace(to_replace='否',  value=0, inplace=True)
df_member['性别'].replace(to_replace='女', value=0, inplace=True)
df_member['性别'].replace(to_replace='男', value=1, inplace=True)
```

而对于各种会员套餐类型，我们把汉字“是”、“否”转换成易于机器读取的布尔类型变量。

复制代码

```
# 其它的是、否字段转换成布尔型数据
binary_features = ['玫瑰套餐', '紫罗兰套餐', '郁金香套餐',
                   '百合套餐', '康乃馨套餐', '胡姬花套餐',
                   '生日套餐','情人节套餐']
for field in binary_features:
    df_member[field] = df_member[field] == '是'
```

这时候，数据集情况如下：

	用户码	性别	玫瑰套餐	紫罗兰套餐	郁金香套餐	百合套餐	康乃馨套餐	胡姬花套餐	生日套餐	情人节套餐	会员卡类型	入会月数	会费支付方式	平均月消费	总消费	已停付会费
0	1	0	True	True	False	True	False	True	True	False	年卡	9	手工转账	65.60	593.30	0
1	2	1	False	False	False	False	False	False	False	True	月卡	9	手工转账	59.90	542.40	0
2	3	1	False	False	False	False	True	False	False	False	月卡	4	就餐时付费	73.90	280.85	1
3	4	1	True	False	False	True	True	False	True	True	月卡	13	就餐时付费	98.00	1237.85	1
4	5	0	False	True	False	False	False	True	True	True	月卡	9	花呗付款	69.40	571.45	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	7039	0	False	False	True	False	False	True	False	False	年卡	13	手工转账	55.15	742.90	0
7039	7040	1	False	False	False	True	False	False	False	False	月卡	2	手工转账	50.30	92.75	0
7040	7041	1	True	False	False	False	False	False	False	True	月卡	22	就餐时付费	85.10	1873.70	1
7041	7042	1	True	True	True	False	True	True	False	True	双年卡	67	手工转账	67.85	4627.65	0
7042	7043	1	True	True	True	True	True	False	True	True	双年卡	63	就餐时付费	59.00	3707.60	0

7043 rows × 16 columns

下一步，我们要做另外一个可视化工作，就是显示会员的留存曲线，这个曲线在留存分析中非常重要，可以让我们发现用户留存相关的很多秘密，所以，我专门拉出来好好给你讲一讲。

Kaplan-Meier 生存模型：显示整体留存曲线

绘制留存曲线，通常的做法是用 Python 手工绘图，这里我要向你推荐一个非常好用的工具包：lifelines（生命线库），它的优点是简单快捷，还能显示整个数据集中用户随着时间推移的留存情况。

lifelines 最早是由精算师和医疗专业人员开发的，通过病人的病情和其他特征，来分析病人的生存可能性。后来，数据分析师们发现，这个模型既然能够判断哪些因子会影响病人的存活率（死亡率），那么，要是把因子换成用户行为特征和产品特征，不就也能分析出用户的留存率（流失率）吗？于是，这个工具被运营人员和数据分析师利用了起来，为分析用户留存提供了很多有用的方法。

在 lifelines 工具包中有一个 Kaplan-Meier 生存模型，它能显示整体的用户留存曲线，通过绘制不同群体的留存曲线，我们就可以观察到数据集中的各个特征对于留存的影响大小。

现在，我们回到这个项目的数据集，观察一下哪些指标与“留存时间”的关系最密切：

	用户码	性别	玫瑰套餐	紫罗兰套餐	郁金香套餐	百合套餐	康乃馨套餐	胡姬花套餐	生日套餐	情人节套餐	会员卡类型	入会月数	会费支付方式	平均月消费	总消费	已停付会费
0	1	0	True	True	False	True	False	True	True	False	年卡	9	手工转账	65.60	593.30	0
1	2	1	False	False	False	False	False	False	False	True	月卡	9	手工转账	59.90	542.40	0
2	3	1	False	False	False	False	True	False	False	False	月卡	4	就餐时付费	73.90	280.85	1
3	4	1	True	False	False	True	True	False	True	True	月卡	13	就餐时付费	98.00	1237.85	1
4	5	0	False	True	False	False	False	True	True	True	月卡	9	花呗付款	69.40	571.45	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	7039	0	False	False	True	False	False	True	False	False	年卡	13	手工转账	55.15	742.90	0
7039	7040	1	False	False	False	True	False	False	False	False	月卡	2	手工转账	50.30	92.75	0
7040	7041	1	True	False	False	False	False	False	False	True	月卡	22	就餐时付费	85.10	1873.70	1
7041	7042	1	True	True	True	False	True	True	False	True	双年卡	67	手工转账	67.85	4627.65	0
7042	7043	1	True	True	True	True	True	False	True	True	双年卡	63	就餐时付费	59.00	3707.60	0

7043 rows × 16 columns

你看出来了吗？它们分别是最后一个字段‘已停付会费’，以及第三个字段‘入会月数’”。有了这两个指标，我们就可以计算出这些用户的生命周期长度，也就是他们大概

在入会多长时间后会流失了。

首先我们要安装 lifelines 工具包：

```
pip install lifelines
```

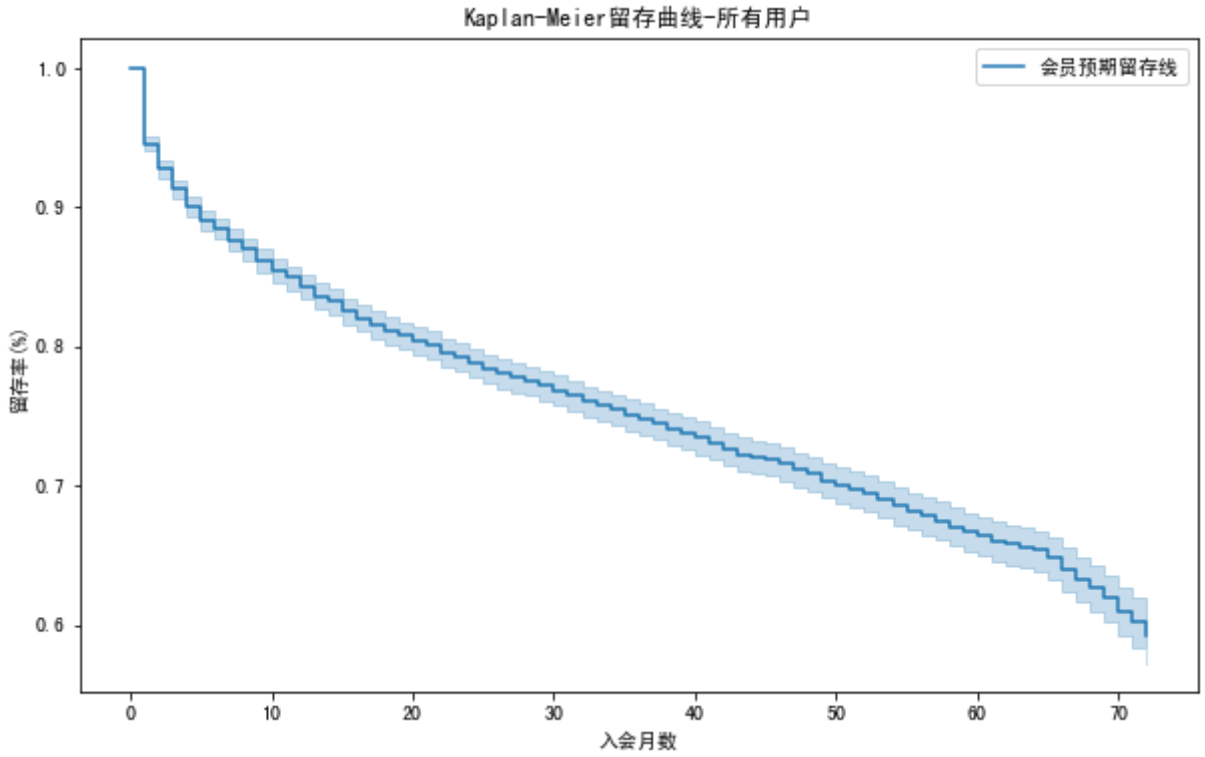
复制代码

下面，我们导入 lifelines，用其中的 Kaplan-Meier 生存模型来查看普通用户随时间而变化的留存率。我们把“入会月数” 和“已停付会费”两个字段输入该模型，这个模型可以拟合用户数据，并以绘制出含置信区间的用户留存曲线。

```
import lifelines #导入生存分析工具包
kmf = lifelines.KaplanMeierFitter() #创建KMF模型
kmf.fit(df_member['入会月数'], #拟合易速鲜花会员流失数据
        event_observed=df_member['已停付会费'],
        label='会员预期留存线')
fig, ax = plt.subplots(figsize=(10,6)) #画布
kmf.plot(ax=ax) #绘图
ax.set_title('Kaplan-Meier留存曲线-易速鲜花会员们') #图题
ax.set_xlabel('入会月数') #X轴标签
ax.set_ylabel('留存率(%)') #Y轴标签
plt.show() #显示图片
```

复制代码

输出如下：



你看，图中的曲线有蓝色的阴影，这个阴影就是置信区间。置信区间是一个范围，比如说有一批学生的考试成绩是从 0 到 100 分，如果我们给出一个分数的区间[55 分 -95 分]，然后说 “该区间的置信水平为 95%” ，这就代表学生的考试成绩落在这个范围的概率是 95%。这里的[55 分 -95 分] 就被称做置信水平为 95% 的学生考试成绩的置信区间。一般来说，我们使用 95% 的置信水平进行区间估计。

从我们这个项目来看，考虑到图中留存率的蓝色阴影所覆盖的范围，我们可以说入会月数为 20 个月之后，有 95% 的可能，用户留存的概率在 78%-82% 之间。而在第 70 个月之后，也就是注册为会员 5 年之后，留存率有 95% 的可能在 58%-64% 这个概率区间。

不过，上面的留存曲线并不能给我们提供更多的用户特定分组的留存情况的对比信息，我们还需要从用户特征入手，分组绘制留存曲线，然后在同一张图中显示，这样可以帮我们挖掘出更多影响用户留存率的因素。

下面，我们就根据用户特征来绘制不同的细分留存曲线。我们先创建一个函数，让这个函数根据用户具体特征，也就是生存分析中的因子，来专门绘制留存曲线：

复制代码

```
def life_by_cat(feature, t='入会月数', event='已停付会费', df=df_member, ax=None)
    for cat in df[feature].unique(): #遍历类别
        idx = df[feature] == cat #当前类别
        kmf = lifelines.KaplanMeierFitter() #创建KaplanMeier模型
```

```
        kmf.fit(df[idx][t], event_observed=df[idx][event], label=cat) #拟合模型
        kmf.plot(ax=ax, label=cat) #绘图
```

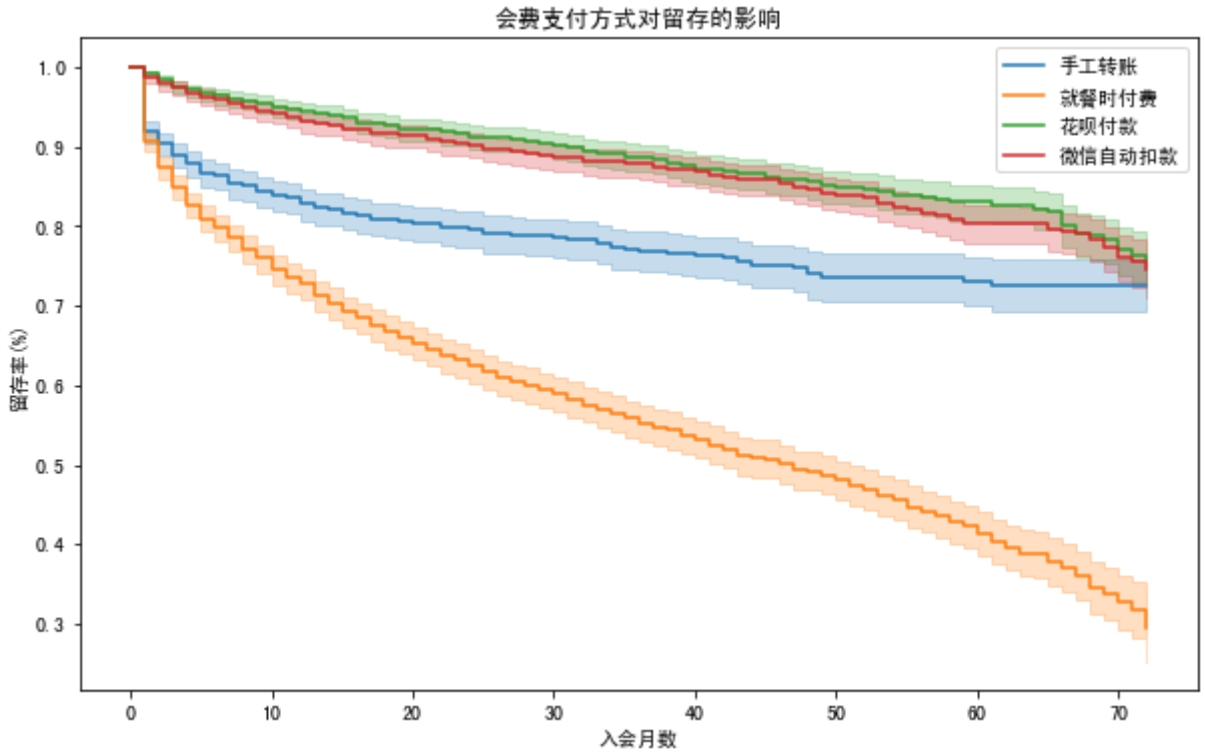
有了这个函数，我们就可以显示用户细分类别的留存曲线，并进行对比。

下面，我们先看看会员费缴纳方式对留存的影响，操作很简单，我们只需要把这个字段传输进刚才定义的 life_by_cat 函数即可：

复制代码

```
fig_pmt, ax_pmt = plt.subplots(figsize=(10,6)) #画布
life_by_cat(feature='会费支付方式', ax=ax_pmt) #调用函数
ax_pmt.set_title('会费支付方式对留存的影响') #图题
ax_pmt.set_xlabel('入会月数') #X轴标签
ax_pmt.set_ylabel('留存率(%)') #Y轴标签
plt.show() #显示图片
```

输出如下：



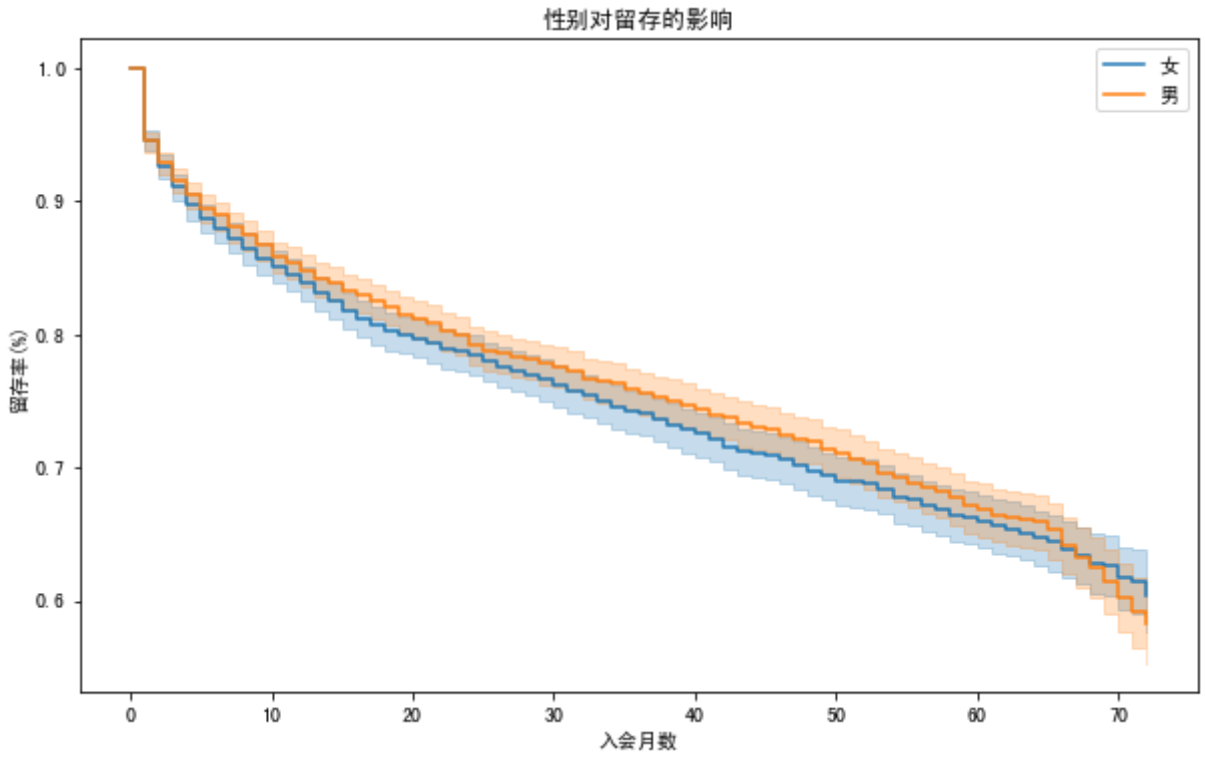
可以看到，如果选择就餐时付会员费，那么流失率将大幅上升。那我们就有理由怀疑，如果要客户下次就餐时支付会员费，那会使很多用户干脆不再前来就餐。另一个对留存有负面影响的支付方式是手工转账支付会员费。而最佳的会员费支付方案是说服用户采用微信自动扣款，或通过花呗付款。

接着，我们再借助这个 life_by_cat 函数，来看看男生与女生的留存周期有没有显著差异。你先猜一猜，性别对于易速鲜花的会员留存率的影响，会不会像付款方式那么大？

复制代码

```
fig_gender, ax_gender = plt.subplots(figsize=(10,6)) #画布
life_by_cat(feature='性别', ax=ax_gender) #调用函数
ax_gender.set_title('性别对留存的影响') #图题
ax_gender.set_xlabel('入会月数') #X轴标签
ax_gender.set_ylabel('留存率(%)') #Y轴标签
plt.show() #显示图片
```

输出如下：



你看，图中两条留存曲线几乎重合，这表明性别差异对易速鲜花用户留存影响并不是十分明显。想想也是，无论男生还是女生，对订购鲜花这件事情的喜好程度好像确实没有什么分别。

到这里呢，我们就通过按特征类别分组的留存曲线，显示出了哪些特征是对用户留存率是有影响的，哪些特征是没有影响的。不过，你可能还是会问：这个留存曲线展示的是全部的用户，或者至少是某一类用户的留存情况，可我的老板就是让我具体关注某几个高价值用户，那么有没有什么方法能展示出某一个人在未来 10 年或者 20 年后的留存情况呢？答案当然是有的。不过，这就要使用 lifelines 库中的另外一个工具，Cox 危害系数模型。

Cox 危害系数模型：预测用户留存概率

Cox 危害系数模型（proportional hazards model）会分析每一个字段对于生存的影响，然后预测出某一个人大概在多久之后会“死亡”，并给出随时间延长，生存的概率值大小，也就是可能性。我们这个场景中的“死亡”，也就是指会员的流失。

所谓“危害”，你可以理解为风险因子，它用来表示分析该特征（或称为因子）是否会增加或减少生存机会，在这里，当然就是指对应特征对用户流失的影响之大小了。然后对用户个体就可以根据这些字段作为特征进行预测分析，判断未来用户留存或者流失的概率。

由于 Cox 危害系数模型可读取的格式是虚拟变量编码，而现在各字段的数据格式是分类编码，在调用该方法之前，我们还需要进一步做点数据的整理工作，把每个字段都规范成为 Cox 模型可以读取的格式。

1. 数据预处理

先通过 Pandas 中的 pd.get_dummies 方法，可以把多类别字段分解为多个二元类别字段，也就是之前在第 8 讲说过的虚拟变量。比如把“会员卡类型”字段，拆分为“会员卡类型 - 年卡”、“会员卡类型 - 月卡”、“会员卡类型 - 双年卡”等。这样一来，输入 Cox 危害系数模型的所有字段都是二元类别字段，都只有 1，0（是，否）两个类别。

你可能觉得奇怪，为什么这个步骤没在之前的特征工程部分完成呢？这是因为我们希望保留“会费支付方式”、“会员卡类型”等这些分类字段来展示不同用户分组的留存曲线，而一旦划分为虚拟变量，原始字段就会消失，所以，我们现在才做这个步骤。

具体的拆解代码如下：

复制代码

```
#把分类字段转换为哑编码
category_features = ['会员卡类型', '会费支付方式'] #要转换的分类的字段
df_member = pd.get_dummies(df_member, #哑变量的个数会比总类别数少1
                           drop_first=True, #这是因为drop_first=True
                           columns=category_features) #这能够避免回归中的多重共线性
df_member #显示数据
```

在这段代码中，关于 drop_first = True 的含义，我们在第 8 讲中已经探讨过了。如果我们为分类变量的 k 个级别保留 k 个虚拟变量，就会存在一个级别的冗余。对于多元线性回归模型来说，这会导致数据集中的多重共线性问题，从而影响回归曲线的建模。

把多类别字段分解为多个二元类别字段后，新的数据集如下：

	用户码	性别	玫瑰套餐	紫罗兰套餐	郁金香套餐	百合套餐	康乃馨套餐	胡姬花套餐	生日套餐	情人节套餐	入会月数	平均月消费	总消费	已停付会费	会员卡类型_年卡	会员卡类型_月卡	会费支付方式_微信自动扣款	会费支付方式_手工转账	会费支付方式_花呗付款
0	1	0	True	True	False	True	False	True	True	False	9	65.60	593.30	0	1	0	0	1	0
1	2	1	False	False	False	False	False	False	False	True	9	59.90	542.40	0	0	1	0	1	0
2	3	1	False	False	False	False	True	False	False	False	4	73.90	280.85	1	0	1	0	0	0
3	4	1	True	False	False	True	True	False	True	True	13	98.00	1237.85	1	0	1	0	0	0
4	5	0	False	True	False	False	False	True	True	True	9	69.40	571.45	0	0	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	7039	0	False	False	True	False	False	True	False	False	13	55.15	742.90	0	1	0	0	1	0
7039	7040	1	False	False	False	True	False	False	False	False	2	50.30	92.75	0	0	1	0	1	0
7040	7041	1	True	False	False	False	False	False	False	True	22	85.10	1873.70	1	0	1	0	0	0
7041	7042	1	True	True	True	False	True	True	False	True	67	67.85	4627.65	0	0	0	0	1	0
7042	7043	1	True	True	True	True	True	False	True	True	63	59.00	3707.60	0	0	0	0	0	0

7043 rows × 19 columns

整理好数据后，我们开始创建模型。

2. 创建并拟合模型

下面的语句就创建出 Cox 模型。

复制代码

```
cph = lifelines.CoxPHFitter() #创建CoxPH模型
```

然后，我们用数据集拟合模型：

复制代码

```
cph.fit(df_member, duration_col='入会月数', #拟合模型
        event_col='已停付会费', show_progress=False)
```

模型拟合结束后，就可以显示某一个用户的生存周期了。

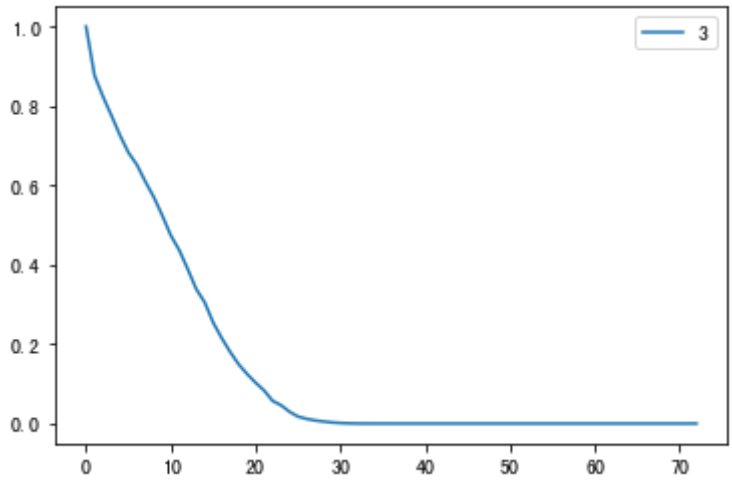
3. 显示留存概率曲线

我们通过模型中的 predict_survival_function 函数，来显示对其中某一个用户的生存概率的预测：

复制代码

```
#显示会员生存曲线
cph.predict_survival_function(df_member.loc[3]).plot()
```

留存概率曲线输出如下：



对这个用户来说，初始的留存概率当然是 1.0；在注册会员 10 个月之后，他留存下来的可能性降低到 50% 左右；在注册会员 20 个月之后，他留存的概率减少到了 20% 以下；30 个月之后，这个用户就极有可能会流失，此时的留存概率几乎为 0。

到这里，我们已经能够预测出用户的留存时间长短，但是怎么知道每一个指标（特征字段），对用户流失的影响有多大呢？其实，Cox 危害系数模型仍然能帮到我们。

Cox 危害系数模型：分析影响留存的因子

Cox 危害系数模型的用途不仅限于预测用户留存的概率，它还可以用来挖掘各个特征字段和用户留存（也就是生命周期）的关联程度。

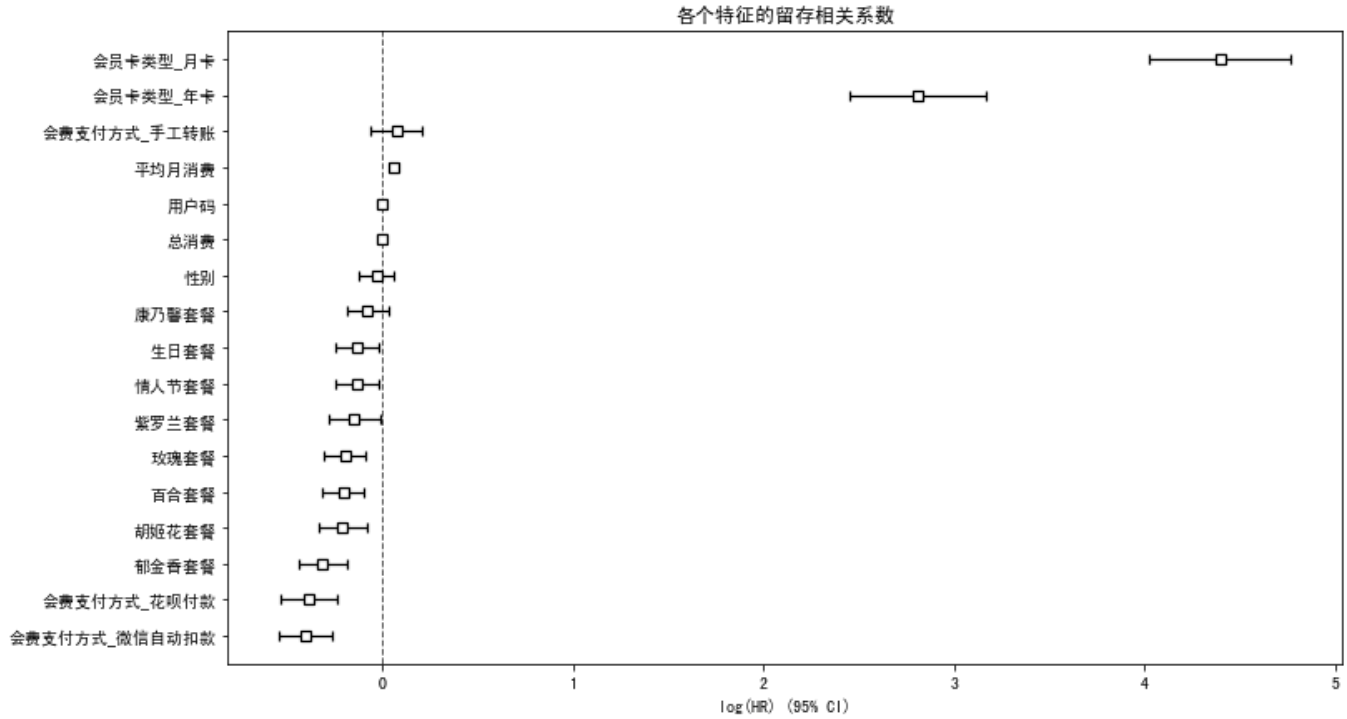
下面这段代码就通过可视化的方式，来显示每一个特征的危害系数及其置信区间。

复制代码

```
fig, ax = plt.subplots(figsize=(12,7)) #画布
ax.set_title('各个特征的留存相关系数') #图题
```

```
3 cph.plot(ax=ax) #绘图
```

输出如下：



在这个图中，大于 0 的系数值预示着对留存有负面影响，值越大，影响越大；而小于 0 的系数值则对留存有正面影响，值越小，影响越大。

可以看出，像月卡、手工转账支付会员费这些，都是潜在的流失原因。这说明，“易速鲜花”应该努力让用户办长期卡，用自动付费取代手工转账。而各种优惠套餐的危害系数都小于 0，说明参与这些活动的会员留存率会有显著提高，所以，应该让会员积极参与这些活动，有利于留存。

好啦，到这里，我们就成功找到了与用户留存高度相关的因子。那么接下来，运营团队会怎么设计新的会员配套、怎么说服会员尽量签订长期合同、怎么优化付款方式……就都是他们的事儿啦。

总结一下

现在，我们来做个总结。今天我们主要学习了 lifelines 这个生存分析库，这个包中有很多用来进行留存分析的实用工具。

其中，Kaplan-Meier 生存模型能帮助我们显示整体的用户留存曲线，通过绘制不同特征类型的人群的留存曲线，我们可以观察到不同特征对于留存的影响。经过观察，我们发现“易速鲜花”会员入会 5 年之后，整体留存率还能够在 60% 左右。而性别对于留存率的影响并不明显，但是会员缴纳会费的方式，则能够对留存率产生影响，这就指导我们在营销时，应该尽量说服会员在办卡时选择自动付费。

此外，我们还介绍了 lifelines 中的 Cox 危害系数模型，这个模型能帮我们预测特定用户的留存概率，并显示出数据集各个特征与留存相关的系数。这个相关系数和留存曲线一样，让我们能够发现对用户留存率产生正面和负面影响的因子。比如说给会员的各种优惠套餐配套，都对老会员的留存起到了正面作用，而有些套餐，比如郁金香套餐，胡姬花套餐等，正面作用尤为明显。

总之，这两个工具为我们做会员留存分析，提供了非常有用的指导性信息，希望你能用好它。

思考题

这节课就到这里了，我给你留两个思考题：

1. 显示整体留存曲线时，我选择了会费支付方式和性别这两个字段，查看特征字段对留存是否有影响。那么，在这个数据集中，你还能找到哪个字段，会对留存与否产生比较大的影响呢？请你绘制这个字段的留存曲线。
2. 除了我们这里介绍的 Lifelines 生命线工具包，你还能找到，或者用过哪些工具，能够辅助我们找到影响用户留存的因子呢？

欢迎你在留言区和我分享你的观点，如果你认为这节课的内容有收获，也欢迎把它分享给你的朋友，我们下一讲再见！

极客时间

本讲重点

lifelines 工具库可以辅助我们做运营过程中的留存分析。其中，Kaplan-Meier 生存模型可以用来显示各细分类别的留存曲线；COX 比例危害模型可以预测特定用户的留存概率，显示各因子和留存的相关系数。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 0 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 13 | 深度学习（下）：3招提升神经网络预测准确率

下一篇 15 | 二元分类：怎么预测用户是否流失？从逻辑回归到深度学习

精选留言 (4)

写留言

谦

2021-09-29

佳哥，有个问题不是很理解，文中提到把分类字段转换为哑编码的时候选择drop_first = True,这能够避免回归中的多重共线性问题，那么共线性问题在Cox模型中会有什么影响呢？我理解留存率的计算各个特征应该是独立 ，不应该相互影响，所以不明白为什么要把drop_first设为True。以会员卡类型为例，drop_frist = True之后，双年卡就没有出现在数据里面了。这样后面也看不到双年卡的留存概率曲线和留存相关系数了。

展开

作者回复: 比较好的思考。是有这个局限。不过，你在另外一个回答中也指出了，避免共线性是必须做的。这就像我们男，女，肯定是一列数据，不可能男一列，女一列。所以正确做法是N-1。

那么，没有双年卡怎么办。

1. 在路上读者分析的不错：get_dummies方法drop掉了双年卡的数据，不过剩下两个有很大的负面影响，那么双年卡必然有很大的正面影响。
2. 你可以手工的Drop掉中间不那么重要的"年卡"，保留"月卡"和"双年卡"，那么双年卡的数据也就得以显现。--- get_dummies工具是可以手工指定要Drop哪一个类别的。

我想到这两个方案，其它同学还有好建议吗？

1

那时刻
2021-09-29

会员制的确能够提高留存，但是自动付费这个点，是否存在玩家忘记这个自动扣费点而实际上玩家已经流失呢？

展开

作者回复: 的确的确，所以对于流失的判断还可以考虑引入RFM值等行为因子，查看购物频率等等。尤其是R值，那是评估用户是否流失的好指标。不过不同类型的App，R值阈值不同。微信的话，三天都不可能不用，但是携程旅行，可能半年一年用一次。因此不能一概而论。

1

在路上
2021-09-29

佳哥好，影响用户留存的因素包括：会费支付方式、性别、会员卡类型、优惠套餐和月消费金额。从Cox 危害系数模型对影响留存的因子分析结果来看，会员卡类型影响较大。会员卡类型包括三种，月卡、年卡和双年卡，get_dummies方法drop掉了双年卡的数据，不过剩下两个有很大的负面影响，那么双年卡必然有很大的正面影响。

展开

作者回复: 对，是这样的。

1

谦

2021-09-29

如果把drop_first=True 改成drop_first=False，在拟合的时候会报错，估计是共线性问题引起的，随机删掉会引起共线性问题的两列数据df_member_1 = df_member_1.drop(['会费支付方式_微信自动扣款','会员卡类型_年卡'],axis=1)，拟合正常。查阅了官方文档，确实提到了共线性问题会引起矩阵计算出错：
Convergence halted due to matrix inversion problems: This means that there is high col…

展开

作者回复: 是这样的。不Drop_First，特征集肯定会有冗余信息。在另一个回答中我给出了更多解释。

1